Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Monterrey

Análisis de ciencia de datos (Gpo 201)

Entregable Etapa 1, 2 y 3

Equipo 2 | Integrantes:

Marcos Aquino Garcia	|A00835576
Sergio Alejandro Esparza González	|A01625430
Carlos Alberto Gómez San Pedro	|A01658377
Pedro Soto Juárez	|A00837560
José Francisco Obregón Gaxiola	|A00227502
Arath Mendivil Mora	|A01660670

Profesores:

Felipe Castillo Rendón
Luis Daniel Mendoza Morales

26 de Abril del 2024

Entregable Etapa 1

Objetivos del negocio

El INEGI es el Instituto Nacional de Estadística y Geografía de México. Es una institución pública encargada de recopilar, procesar y difundir información estadística y geográfica sobre el país. Realiza censos de población y vivienda, encuestas económicas y sociales, y produce indicadores que son fundamentales para la toma de decisiones tanto en el ámbito gubernamental como en el privado. Su labor es crucial para entender la realidad demográfica, económica y social de México.

Situación del negocio

Al ser el INEGI una institución pública, como mencionamos anteriormente, todas sus estadísticas, datos y entre otros están publicados para todo el que desee analizar la información. A parte de esto, nos proporcionaron códigos acerca del Directorio Estadístico Nacional de Unidades Económicas (DENUE), donde se ofrecen datos de identificación, tamaño, ubicación y giro de los negocios activos del país, y del Censo de Población y Vivienda, que informa acerca de la dimensión, estructura y distribución de la población, así como características sociodemográficas y culturales, y las características de cada vivienda del país, los cuales nos servirán en extraer y analizar la información que necesitamos de sus bases de datos. El INEGI nos dio de igual manera libertad en elegir qué datos queremos analizar.

Se escogió analizar en tres entidades distintas, que serían Ciudad de México, Sonora y Veracruz, la cantidad de personas mayores de 18 años las cuales tienen educación post básica. Esto nos ayudará a predecir algunas de las causas por las cuales en México no se termina la educación post básica analizando datos de vivienda, salud y entre otros.

Metas del Proyecto de Análisis Predictivo y Criterios de éxito

- **Objetivo del Proyecto**

Crear un modelo que prediga si hay correlación en variables que son viviendas sin transporte,viviendas sin internet y población sin afiliación a un servicio de salud , si tiene que ver , si es un obstáculo para la población mayor de 18 años en México que no termina la educación post básica [7]

- **Problema actual**

  Es el abandono escolar de cierta población en México enfocado a estados de Sonora,CDMX y Veracruz en la preparatoria , queremos saber si variables planteadas tienen que ver con ello , usando datadets del censo de población y vivienda.

- **Hipótesis**

Se espera que factores socioeconómicos, como el acceso a servicios básicos, la situación laboral y el entorno familiar, tengan un impacto significativo en las condiciones de vida de la población mayor de 18 años en México que no completa la educación post básica. Por lo tanto, al recopilar datos demográficos, económicos y sociales de esta población, podemos

construir un modelo predictivo que identifique patrones y relaciones entre estas variables y las condiciones de vida, lo que nos permitirá comprender mejor los determinantes de la calidad de vida de este grupo y desarrollar estrategias efectivas para mejorar sus condiciones [2].

- **Metodología CRISP -DM**
  Conocimiento del negocio

  Criterios de éxito: Crear un modelo predictivo que alcance un óptimo de precisión, capaz de identificar con éxito los factores predictivos más significativos que contribuyen a la educación incompleta y, por tanto, afectan la calidad de vida.
  Riesgos y contingencias: Posibles inexactitudes en los datos del INEGI, problemas de integración de datos, y la necesidad de ajustes metodológicos debido a datos faltantes o atípicos.
  Comprensión de los datos del negocio

  Fuentes de datos: Datos proporcionados por el INEGI Censo de Población y Vivienda, con variables como la población con educación posbásica y características demográficas básicas.
  Técnicas de exploración de datos: Análisis descriptivo utilizando medidas de tendencia central y dispersión, histogramas para análisis de distribución, y boxplots para identificar valores atípicos.
  Calidad de datos: Verificar la completitud y exactitud de los datos, manejar valores faltantes y corregir errores.

  Preparación de los datos
  Selección de datos: Se hará énfasis en las variables como PSINDER (población sin acceso a servicios de salud), VPH_NDACMM (viviendas sin vehículo) y VPH_SINCINT (viviendas sin internet).
  Limpieza de datos:, Manejo de valores faltantes mediante métodos de imputación apropiados, y conversión de categorías a numérico si es necesario.

  Integración de datos: Combina datasets de diferentes entidades para un análisis consolidado.
  Formateo de datos: Asegurar que todos los datos estén en formatos adecuados para el análisis.

  Modelación
  Técnicas de modelado: Emplear algoritmos de predicción para comparar los grupos de interés. Métodos potenciales incluyen K Neighbors, árboles de decisión, o random forests.
  Configuración de pruebas: División de datos en conjuntos de entrenamiento y prueba para validar la efectividad del modelo.

Criterios de evaluación: Utilizar la precisión del modelo, la matriz de confusión, R^2, entre otros indicadores de desempeño.

Evaluación de los modelos

Evaluación de resultados: Comparar los resultados del modelo con los objetivos y criterios de éxito del negocio. Interpretar los modelos en términos de impacto social y económico.
Revisión del proceso: Revisar cada etapa del proceso CRISP-DM para asegurar que se ha ejecutado correctamente y ajustar donde sea necesario.
Pasos a seguir: Decidir sobre la implementación del modelo y la formulación de recomendaciones de política basadas en los hallazgos del modelo.

Implementación de la solución

Implementar el modelo en el entorno de producción en sistemas existentes o desarrollando una interfaz de usuario si es necesario.
Asegurarse de que el modelo esté correctamente configurado y pueda generar predicciones en tiempo real según sea necesario
Establecer un proceso de monitoreo continuo para supervisar el rendimiento del modelo en producción.
Realizar ajustes periódicos al modelo según sea necesario, como la actualización con nuevos datos o la recalibración de parámetros.

- **Resultados Esperados**

Se espera que el modelo proporcione información valiosa y utilizable para mejorar las condiciones de vida de las generaciones que vienen y así tener un enfoque de que son los problemas que se necesitan solucionar para que un niño pueda concluir con su educación así como una comprensión más profunda de los factores que influyen en su calidad de vida.

- **Beneficios Esperados**

El desarrollo de un modelo predictivo para las condiciones de vida de la población mayor de 18 años en México que no completa la educación post básica tiene como objetivo identificar los factores que influyen en su calidad de vida y proporcionar información valiosa para la formulación de políticas y la asignación de recursos. Los resultados esperados incluyen la identificación de necesidades específicas, el desarrollo de políticas efectivas, la optimización de recursos, la reducción de desigualdades y la mejora de la calidad de vida. En resumen, el modelo tiene el potencial de contribuir significativamente a la mejora de las condiciones de vida de este grupo demográfico vulnerable en México. [6]

- **Impacto Social**

El desarrollo de un modelo predictivo para las condiciones de vida de la población mayor de 18 años en México que no completa la educación post básica tiene el potencial de tener un impacto social positivo al reducir la pobreza y la desigualdad, mejorar el bienestar y la calidad de vida, promover la inclusión social y fortalecer la cohesión social. [8]

Plan del proyecto

Etapa	Tiempo	Responsable
Entregable 1: Conocimiento del negocio	2 horas	Marcos Aquino Garcia
Entregable 1 Conocimiento del negocio	2 horas	Sergio Alejandro Esparza González
Entregable 2: Integración, Comprensión y Preparación de los datos	2 horas	Carlos Alberto Gómez San Pedro
Entregable 2: Integración, Comprensión y Preparación de los datos	2 horas	Pedro Soto Juárez
Entregable 3: Presentación	2 horas	José Francisco Obregón Gaxiola
Entregable 3: Presentación	2 horas	Arath Mendivil Mora

Marco Teórico

El abandono escolar es un desafio mundial , como el propuesto en nuestro reto y también en otros entornos tales como la universidad y los MOOCS .La capacidad de predecir un abandono es importante de tal manera para hacer algo al respecto , exploraremos como un random forest en dos estudios se utilizó para predecir el abandono escolar

Antecedentes

Se ha abordado también predecir dichos problemas usando técnicas de aprendizaje automático ,modelos lineales,random forest y decision trees.

Random Forest en MOOCS [4]
En un estudio enfocado a matemáticas en Edx con una universidad estatal de Arizona se utilizó random forest para predecir el abandono escolar se obtuvieron estas métricas:
Precisión: 88%
Recall: 87.5%
F1-score: 87.5%
AUC (Área bajo la curva ROC): 94.5%

Random Forest en Universidades [1]
En otro estudio se utilizó random forest para ver si abandonaban la universidad en Alemania.

Esta métrica indica que el modelo tenía una capacidad significativa para predecir la deserción estudiantil en el contexto universitario.

AUC (Área bajo la curva ROC): 0.86

Los modelos con una buena métrica pueden saber si una persona o no puede abandonar sus estudios.

Anexos

- Marcos Aquino Garcia:

El abandono escolar también tiene género

Resumen: El abandono escolar es más común entre hombres durante la pandemia. Las mujeres suelen abandonar la escuela por falta de recursos o por eventos como el matrimonio o el embarazo, mientras que los hombres lo hacen más por necesidad de trabajar o percepción de falta de habilidades. Aunque las mujeres pueden tener expectativas salariales más altas con más educación, las condiciones laborales favorecen a los hombres, lo que los lleva a necesitar menos educación para obtener ingresos similares. Esto refuerza los estereotipos de género y destaca la necesidad de políticas que aborden estas inequidades.

- Arath Mendivil Mora:

Boletín 100 Desciende a 8.1% tasa de abandono escolar en Educación Superior

Resumen: El artículo publicado por la Secretaría de Educación Pública (SEP) de México

reporta un descenso significativo del 8.1% en la tasa de abandono escolar en el nivel de educación superior, destacando los esfuerzos del Gobierno de México para mejorar la continuidad educativa en este nivel. Luciano Concheiro Bórquez, subsecretario de Educación Superior, ha enfatizado que este logro es resultado de diversas medidas implementadas para fortalecer el sector, incluyendo la mejora de infraestructuras y la provisión de becas y apoyos financieros, especialmente en el contexto de los desafíos impuestos por la pandemia de COVID-19. La SEP se compromete a seguir trabajando en reducir aún más el abandono escolar y mejorar la calidad de la educación superior en el país.

- José Francisco Obregon Gaxiola:

Abandono escolar de niñas, niños y adolescentes en México (2016-2022)

Resumen: El texto proporciona datos y análisis sobre el abandono escolar de niñas, niños y adolescentes en México entre los años 2016 y 2021. Se destaca que, según datos de la Medición de la Pobreza y de la SEP, millones de niños y adolescentes abandonan la escuela, lo que representa un desafío significativo para garantizar su derecho a la educación. Se señala que las tasas de abandono escolar son más altas entre los hombres que entre las mujeres en todos los niveles educativos. Además, de que se identifican regiones específicas del país donde el abandono escolar es más pronunciado. El texto también ofrece recomendaciones para abordar este problema, incluida la necesidad de políticas que promuevan la inclusión y la accesibilidad educativa.

- Carlos Alberto Gómez San Pedro:

School Location, School Section and Students' Gender as Predictors to Secondary School Dropout Rate in Rivers State, Nigeria

Resumen: Este estudio se realizó para subrayar en qué medida las variables de ubicación de la escuela, género del estudiante y sección escolar pueden predecir la tasa de deserción de los estudiantes de secundaria. Se adoptó un diseño ex post facto y todos los datos sobre inscripción, retención y finalización de los estudiantes se recogieron de los registros escolares disponibles para dos cohortes de estudiantes y abarcando cuatro sesiones, utilizando una hoja de recopilación de datos de deserción escolar desarrollada por el investigador (ioonSDDCS, por sus siglas en inglés). El resultado de los datos analizados muestra tasas de deserción de aproximadamente el 19% con respecto a las variables predictoras.

- Pedro Soto Juárez:

Causas y consecuencias de la deserción escolar en el bachillerato:Caso universidad autónoma de sinaloa

Resumen: Una de las causas que más influyen en la deserción de los estudios de preparatoria son: casarse y el reprobar materias , donde se destaca en este estudio que las mujeres desertan más que los hombres debido a que quieren salir de su cotidianidad y buscar otras oportunidades al lado de su novio sin embargo esto causa que tengan que atender una familia a una temprana edad y se les priva de un desarrollo profesional , la segunda causa es el que no estan motivados los estudiantes por lo tanto no les gusta estudiar , el factor económico de que sus padres no ganan de una manera que puedan seguir pagandoles los estudios y la falta de apoyo de los padres que estudien.

- Sergio Alejandro Esparza González:

Los centros escolares y su contribución a paliar el desenganche y abandono escolar

El tema del desenganche y el abandono escolar es un tema de diversas aportaciones teóricas e investigaciones recientes que toman como foco de atención el centro escolar y lo que acontece en él. La estrecha relación que existe entre el desapego o la no-implicación del alumno en su formación escolar y la experiencia educativa que brinda el centro escolar son factores importantes para el abandono escolar. Existen dos planteamientos principales acerca de los aspectos pedagógicos y organizativo de centros educativos que enganchan a los alumnos, acerca de la eficacia escolar y los análisis críticos y socio-críticos; desde ambas perspectivas se comentan propuestas de actuación, referidas a las relaciones y clima relacional de la enseñanza, y a las relaciones con la familia y el entorno.

Etapa 2

Dimensión del dataset:

Unificamos los datos de 3 Estados para tenerlo todo junto en un mismo dataset, los resultados de las dimensiones fueron los siguientes:

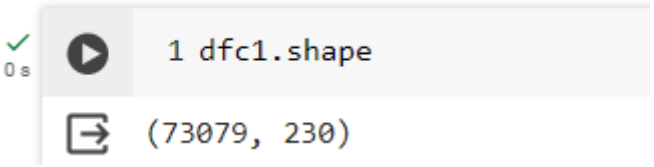


Figura 1. Dimensión del dataset.

Se tienen 237,913 filas con 230 columnas.

Describe claramente en una hoja adicional del documento en Excel cada uno de los datos, incluyendo su nombre, descripción, tipo (categórico/Numérico), valores posibles que puede tomar, y total de valores nulos.

ENTIDAD: Número correspondiente a la entidad federativa. Es cualitativo. Los valores posibles que puede tomar es 9, 26, 30, debido a que solo se seleccionaron tres entidades.
MUN: Número correspondiente al municipio. Es cualitativo y son números enteros positivos
NOM_MUN: Nombre del municipio. Es cualitativo.
LOC: Número de localidad. Es cualitativo. Valores enteros positivos.
NOM_LOC: El nombre de la localidad, es cualitativo.
AGEB: Clave que identifica el interior de una localidad. Se compone de números y letras y es cualitativo
MZA: Número que identifica la manzana, es cualitativo y son números enteros positivos.
P18YM_PB: Población de 18 años y más con educación posbásica. Es numérico. Puede tomar cualquier valor entero arriba de 0.
0.
P_18YMAS: Población de 18 años y más. Es numérico. Puede tomar cualquier valor entero arriba de 0.
PSINDER: Población sin afiliación a servicios de salud. Es numérico. Puede tomar cualquier valor entero de 0 a infinito.
VPH_NDACMM: Viviendas particulares que no disponen de automóvil o camioneta, de motocicleta o motoneta. Puede tomar cualquier valor entero. Puede tomar valor de 0 a infinito.
VPH_SINCINT: Viviendas particulares habitadas sin computadora ni Internet. Puede tomar cualquier valor entero. Puede tomar valor de 0 a infinito.

Estos son los valores nulos de cada columna:

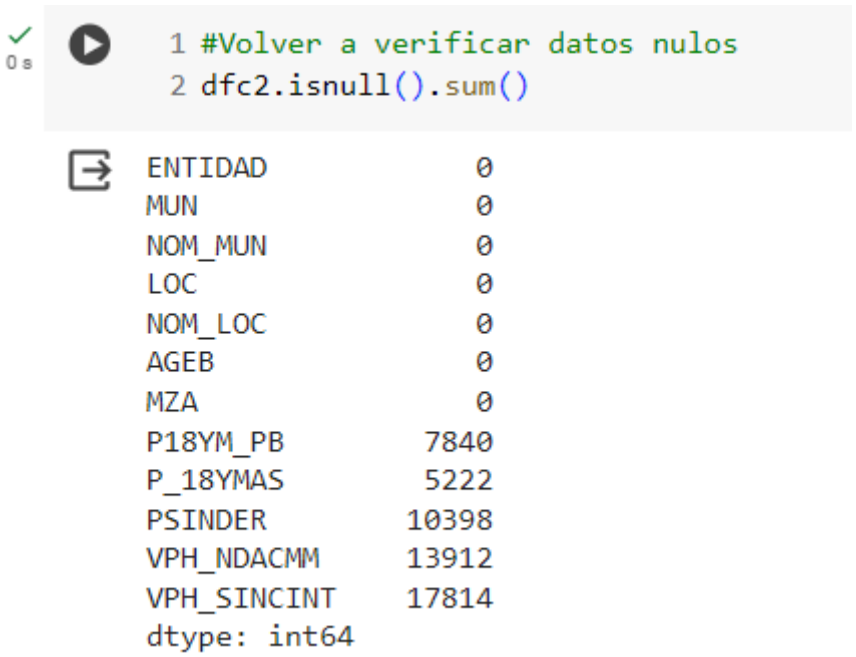


Figura 2. Número de valores nulos por columna.

El rango de valores (mínimo y máximo) que toma cada variable en el dataset se muestra en la figura 3.

Calcula medidas estadísticas

Medidas de tendencia central: promedio, media, mediana y moda de los datos.
Medidas de dispersión: rango: máximo - mínimo, varianza, desviación estándar.

```
[29] #Usamos describe para calcular los valores estadisticos de las variables cuantitativas
     df_cuan.describe(include="all")
```

	P18YM_PB	P_18YMAS	PSINDER	VPH_NDACMM	VPH_SINCINT
count	6.114300e+04	6.114300e+04	6.114300e+04	6.114300e+04	6.114300e+04
mean	3.329617e+02	6.096647e+02	2.115388e+02	1.257298e+02	8.061950e+01
std	2.186200e+04	3.953997e+04	1.435611e+04	8.616970e+03	6.058248e+03
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	9.000000e+00	2.300000e+01	6.000000e+00	4.000000e+00	3.000000e+00
50%	2.500000e+01	5.100000e+01	1.500000e+01	1.000000e+01	7.000000e+00
75%	5.700000e+01	9.900000e+01	3.500000e+01	2.100000e+01	1.300000e+01
max	4.516388e+06	7.166737e+06	2.502789e+06	1.496019e+06	1.329356e+06

Figura 3. Resumen de estadística descriptiva de las variables del dataset.

En esta figura se presentan los valores de la media (mean), mediana (50%), desviación estándar (std), mínimo (min) y máximo (max).

	P18YM_PB	P_18YMAS	PSINDER	VPH_NDACMM	VPH_SINCINT
0	0.0	0.0	0.0	0.0	0.0

Figura 3. Moda de las variables cuantitativas.

En la figura 3 se muestra la moda de cada variable cuantitativa del dataset.

```
0 s
1 range = df_cuan.max()-df_cuan.min()
2 range
```

```
P18YM_PB       4516388.0
P_18YMAS       7166737.0
PSINDER        2502789.0
VPH_NDACMM     1496019.0
VPH_SINCINT    1329356.0
dtype: float64
```

Figura 4. Rango de las variables cuantitativas.

En la figura 4 se muestra el rango de cada variable cuantitativa del dataset.

```
1 df_cuan.var()
```

```
P18YM_PB        5.821137e+08
P_18YMAS        1.905391e+09
PSINDER         2.482453e+08
VPH_NDACMM      9.136296e+07
VPH_SINCINT     4.618711e+07
dtype: float64
```

Figura 5. Varianza de las variables cuantitativas.

En la figura 5 se muestra la varianza de cada variable cuantitativa del dataset.

Variables cualitativas o categóricas

Tabla de distribución de frecuencia
Moda

```
pd.DataFrame(dfStatesQualitative.value_counts())
```

ENTIDAD	count
30	72950
9	65098
26	49732

Figura 6. Tabla de distribución de frecuencias de las variables cualitativas.

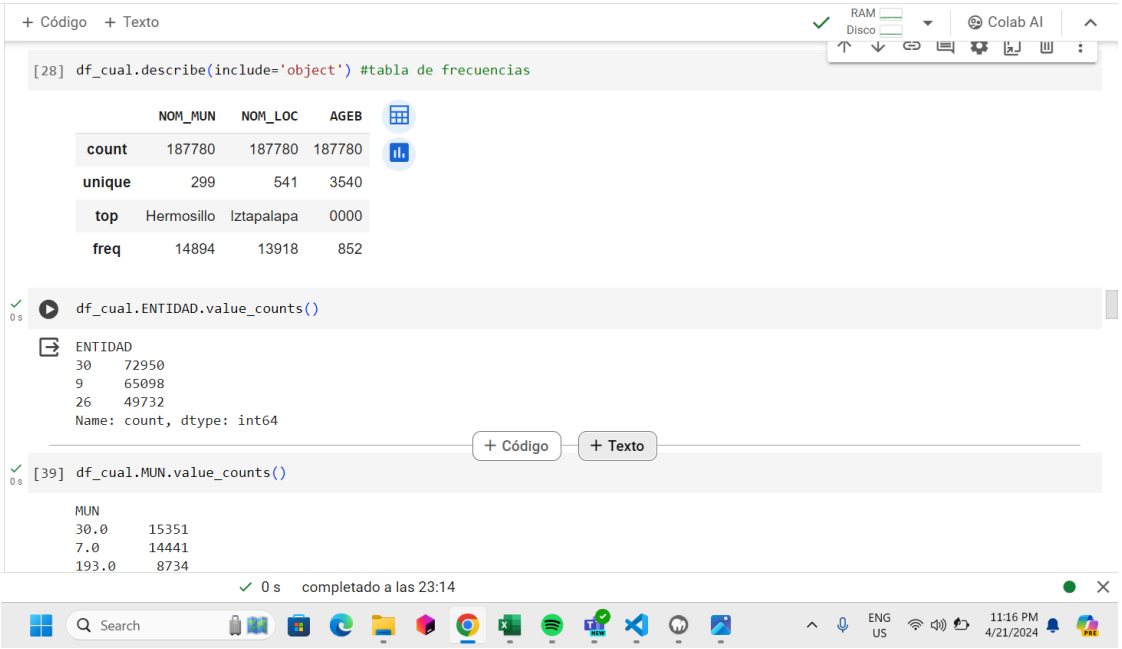
En la figura 6 se muestran las frecuencias de las variables cualitativas.

```
df_cual.mode() #moda de cualitativas
```

	ENTIDAD	MUN	NOM_MUN	LOC	NOM_LOC	AGEB	MZA
0	30	30.0	Hermosillo	1.0	Iztapalapa	0000	0.0

Figura 7. Moda de las variables cualitativas.

De igual manera, en la figura 7, se muestra la moda de las mismas.



```
[28] df_cual.describe(include='object') #tabla de frecuencias
```

	NOM_MUN	NOM_LOC	AGEB
count	187780	187780	187780
unique	299	541	3540
top	Hermosillo	Iztapalapa	0000
freq	14894	13918	852

```
df_cual.ENTIDAD.value_counts()

ENTIDAD
30    72950
9     65098
26    49732
Name: count, dtype: int64

[39] df_cual.MUN.value_counts()

MUN
30.0    15351
7.0     14441
193.0    8734
```


En esta imagen se presenta la tabla de frecuencias de los valores categóricos.

2) Explora los datos usando herramientas de visualización

Variables cuantitativas:

Medidas de posición no-central: cuartiles, outlier (valores atípicos), boxplots

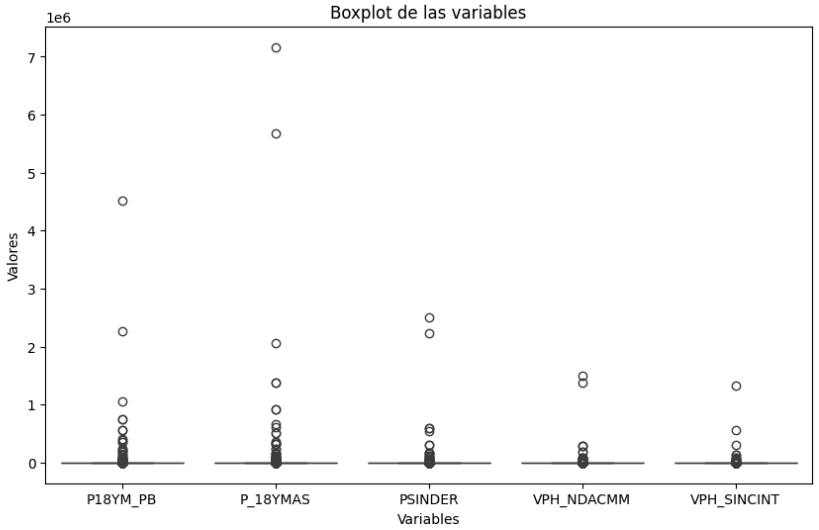


Figura 8. Diagramas de cajas de cada variable cuantitativa.

Se puede observar que en todas las variables existen datos atípicos (muy pocos abajo del bigote de abajo, pero muchos arriba del bigote de arriba). Los cuartiles se pueden observar en

la figura 2.

Análisis de distribución de los datos (Histogramas). Identificar si tiene forma simétrica o asimétrica

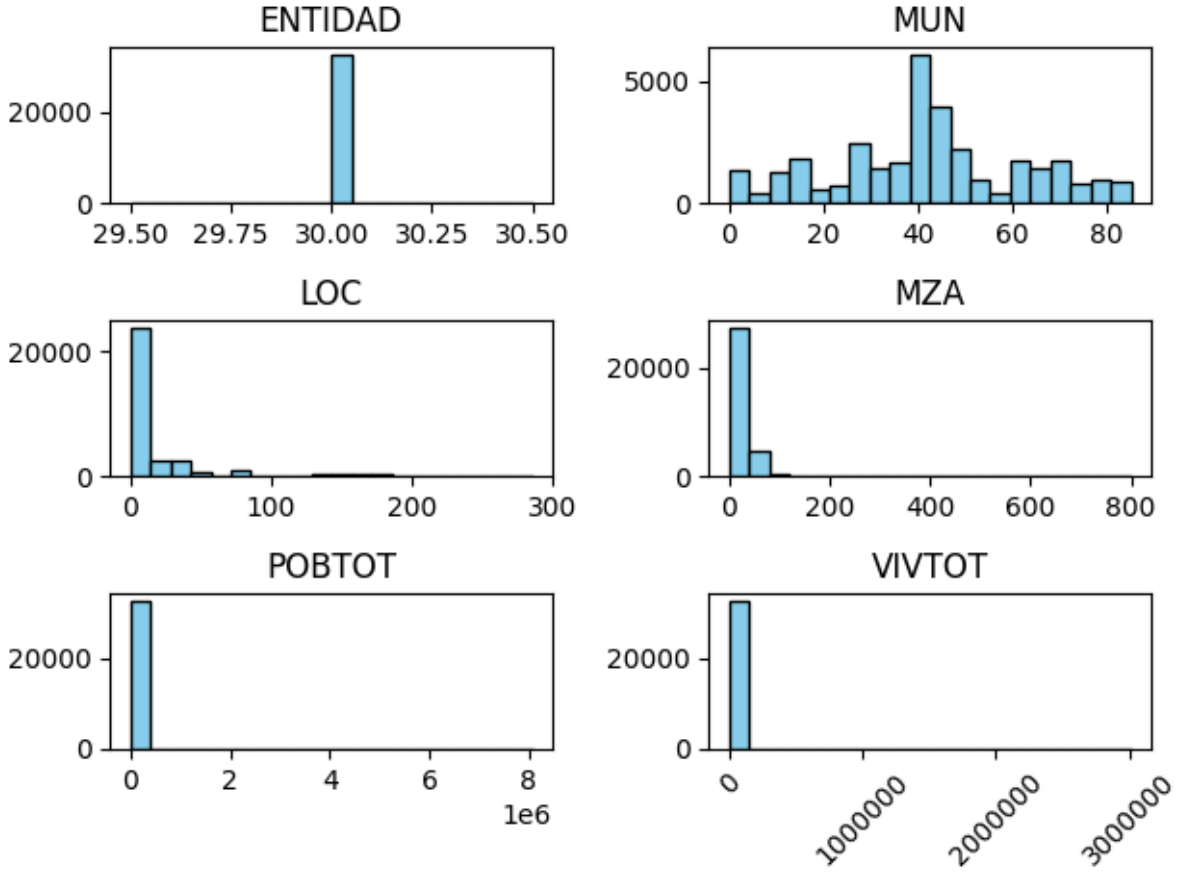


Figura 9. Histogramas de frecuencias de las variables cuantitativas.

Tiene una asimetría positiva porque tiene un sesgo hacia la derecha, pues en esa región les faltan datos, debido a los mismos datos atípicos.

Análisis de correlación de los datos, mapa de color

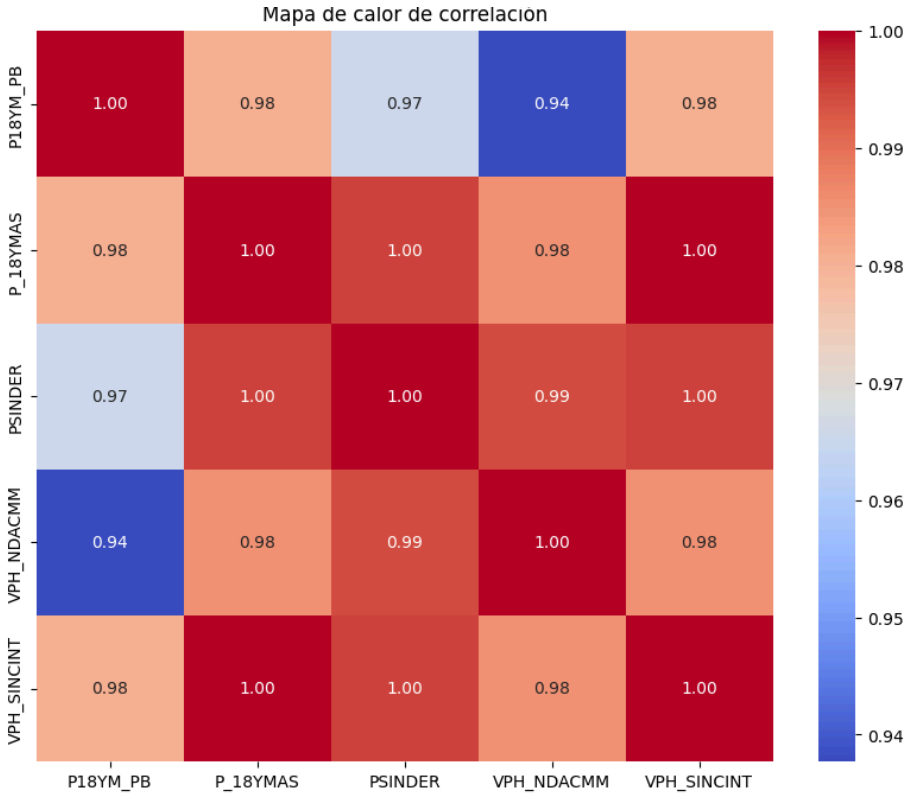


Figura 10. Histogramas de frecuencias de las variables cuantitativas.

En el mapa de calor se puede ver como todas las variables tienen una alta correlación positiva, eso lo consideramos como hallazgo principal antes de usar los modelos de regresión.

Variables cualitativas o categóricas

Distribución de los datos (diagramas de barras, diagramas de pastel)

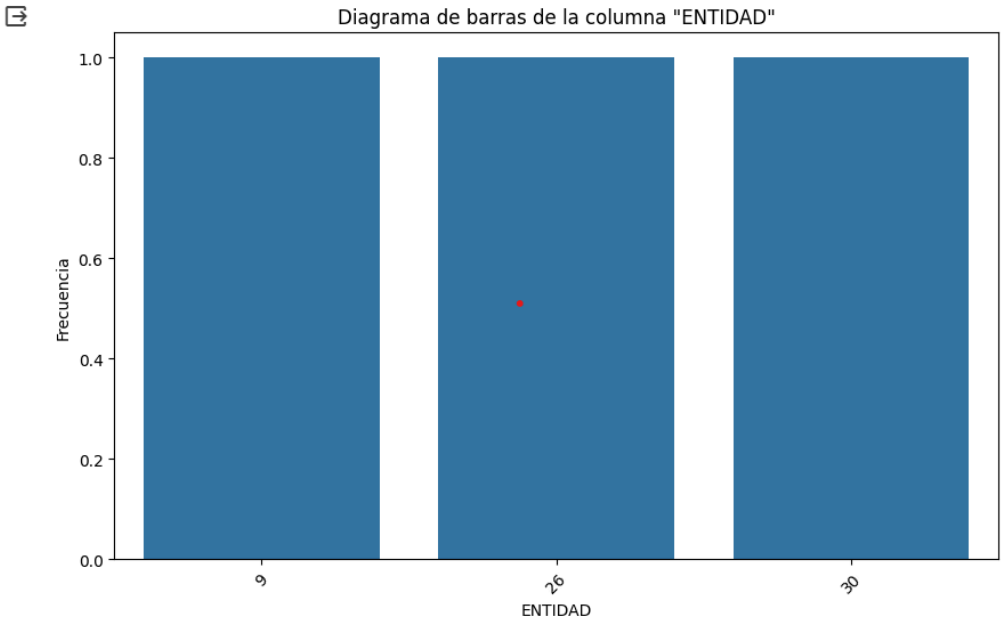


Figura 11. Diagrama de barras de la columna ‘ENTIDAD’.

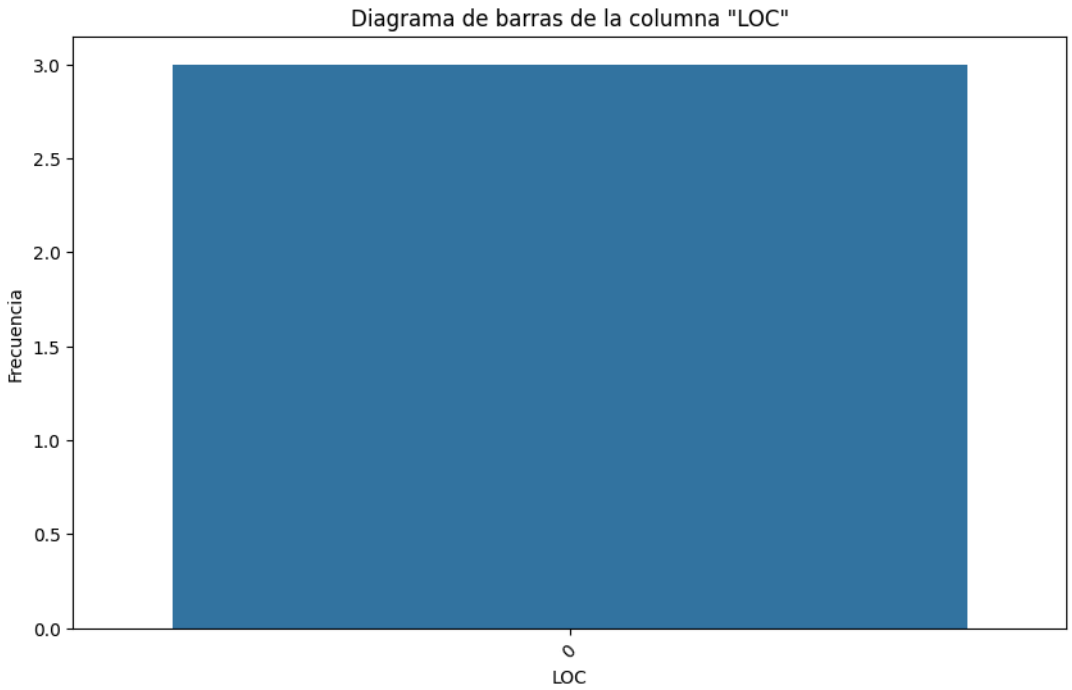


Figura 12. Diagrama de barras de la columna ‘LOC’.

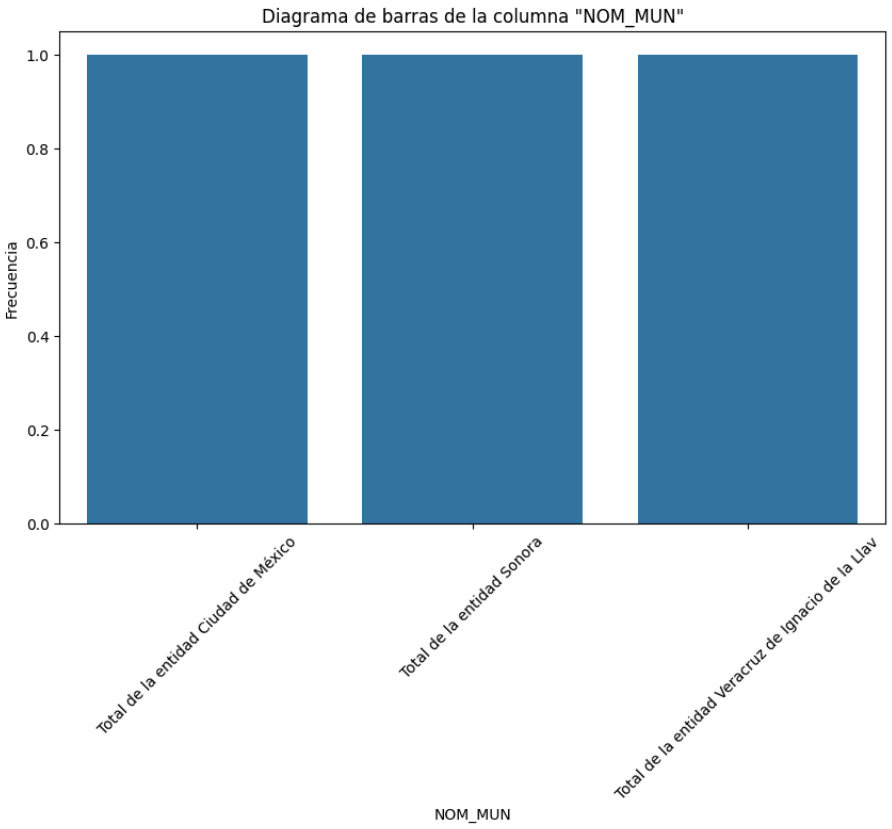


Figura 13. Diagrama de barras de la columna ‘NOM_MUN’.

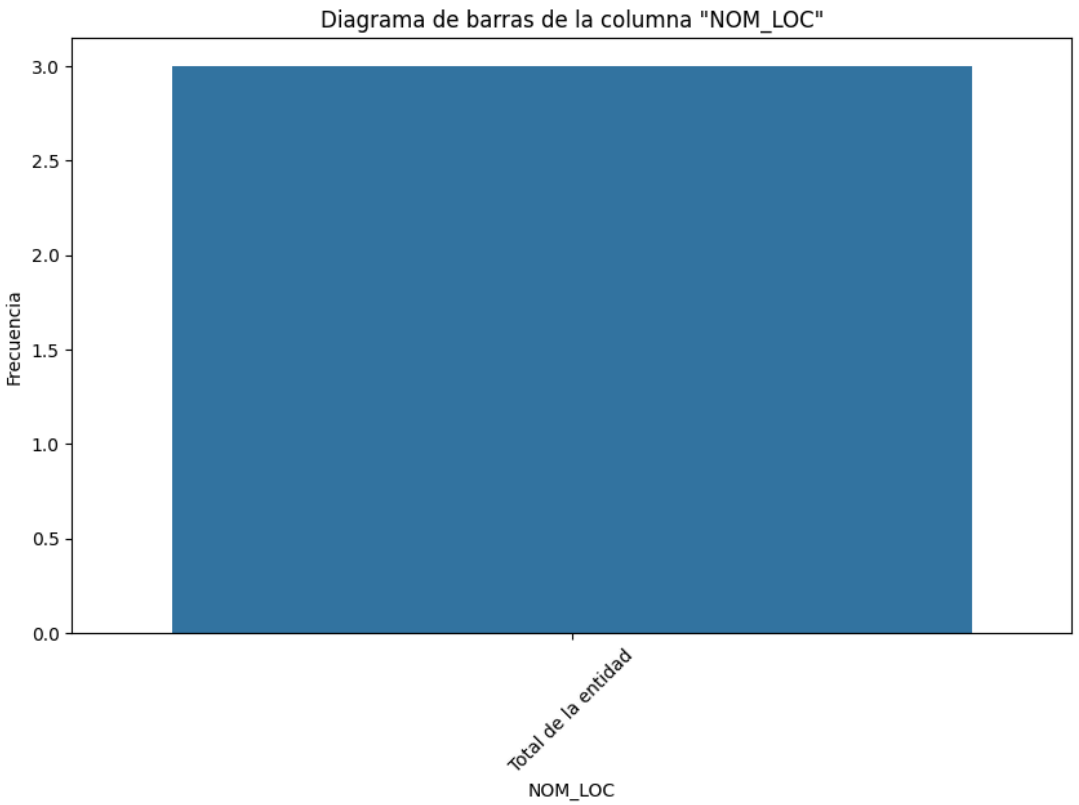


Figura 14. Diagrama de barras de la columna ‘NOM_LOC’.

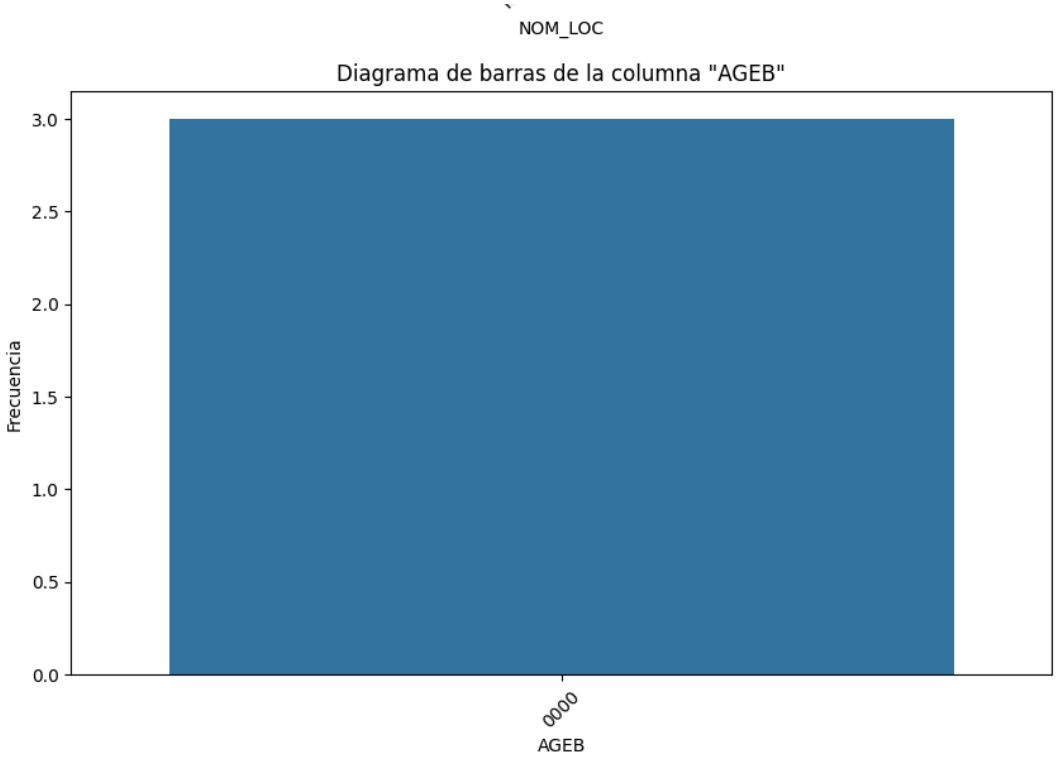


Figura 15. Diagrama de barras de la columna ‘AGEB’.

Figura 16. Diagrama de barras de la columna ‘MZA’.

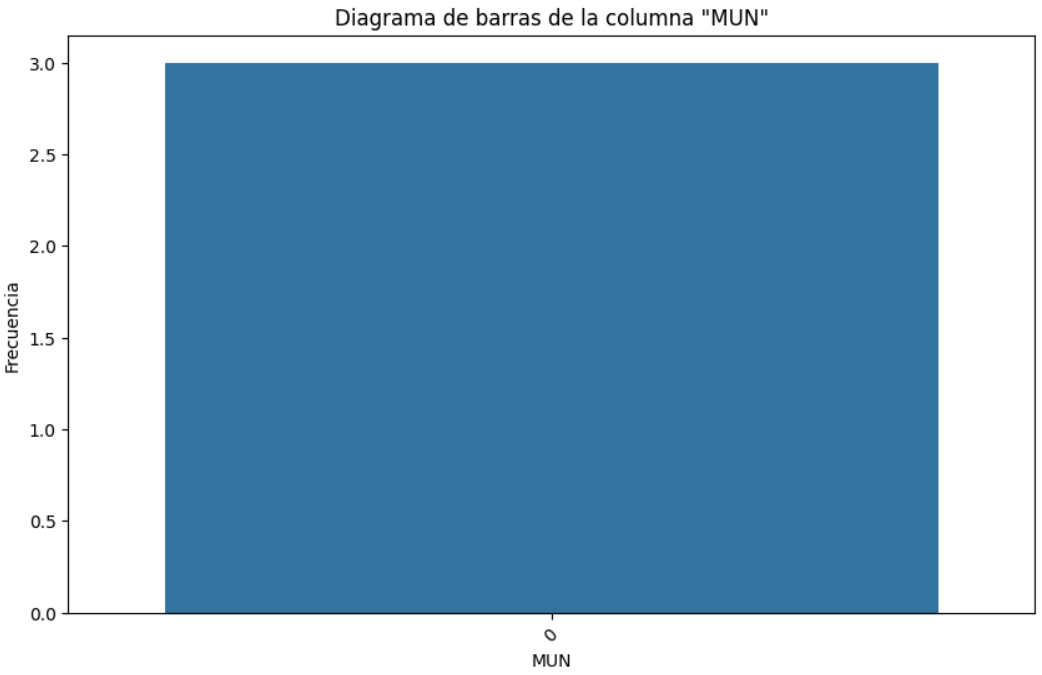


Figura 17. Diagrama de barras de la columna ‘MUN’.

Figura 18. Gráfico de pastel de la columna ‘ENTIDAD’.

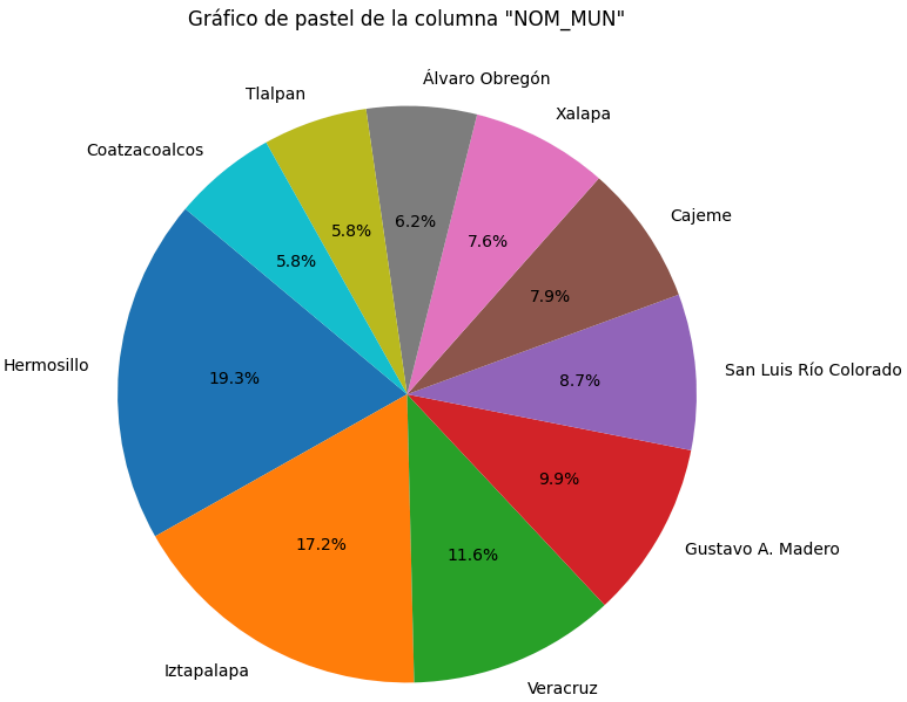


Figura 19. Gráfico de pastel de la columna ‘NOM_MUN’.

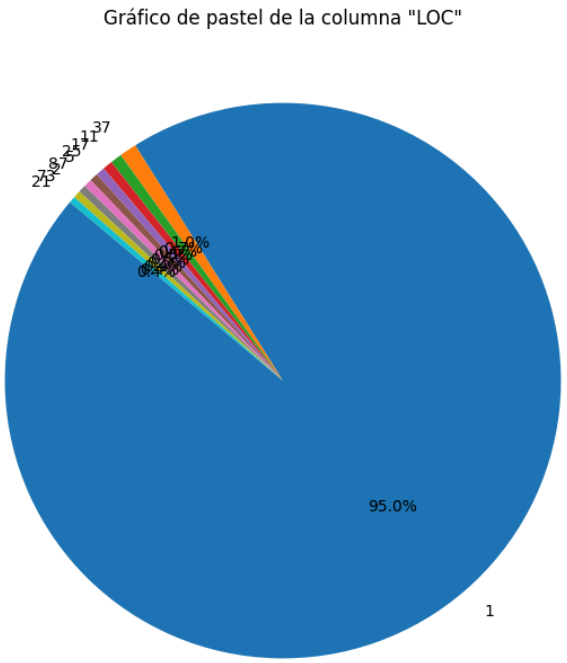


Figura 20. Gráfico de pastel de la columna ‘LOC’.

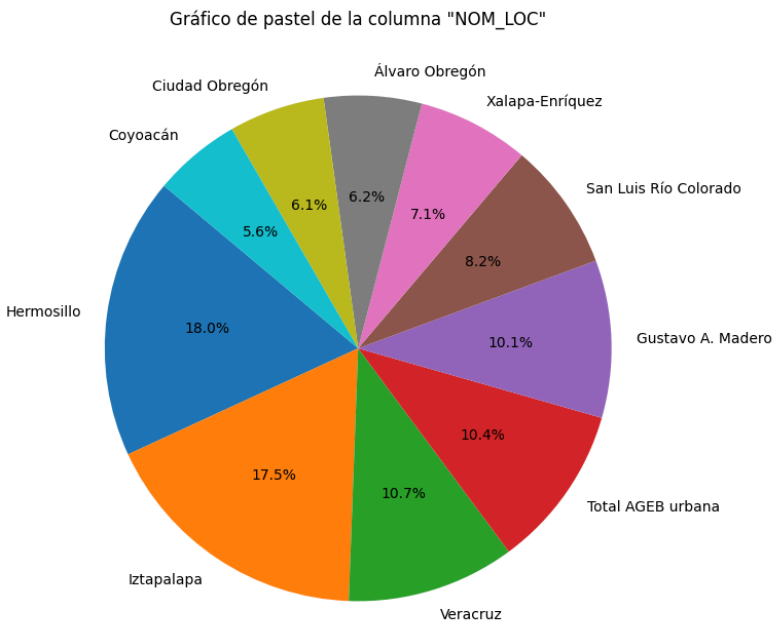


Figura 21. Gráfico de pastel de la columna ‘NUM_LOC’.

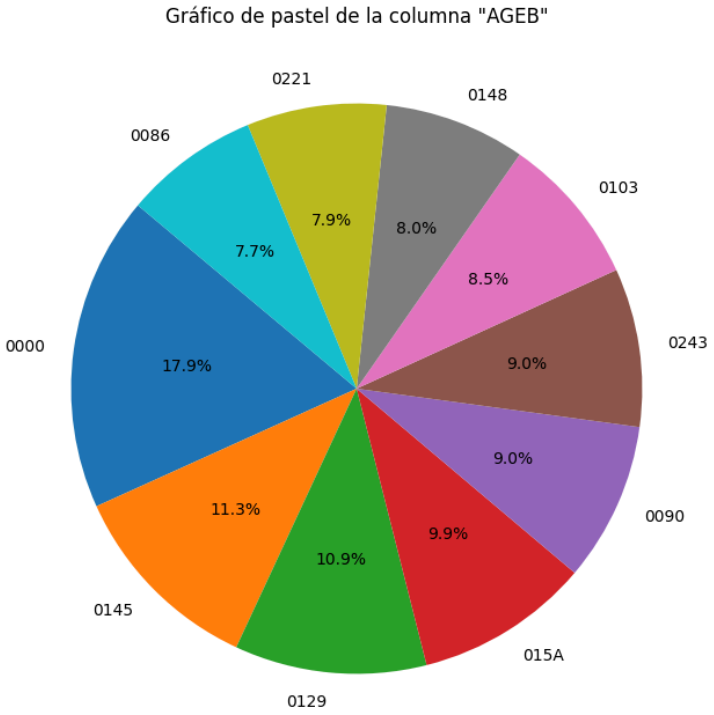


Figura 22. Gráfico de pastel de la columna ‘AGEB’.

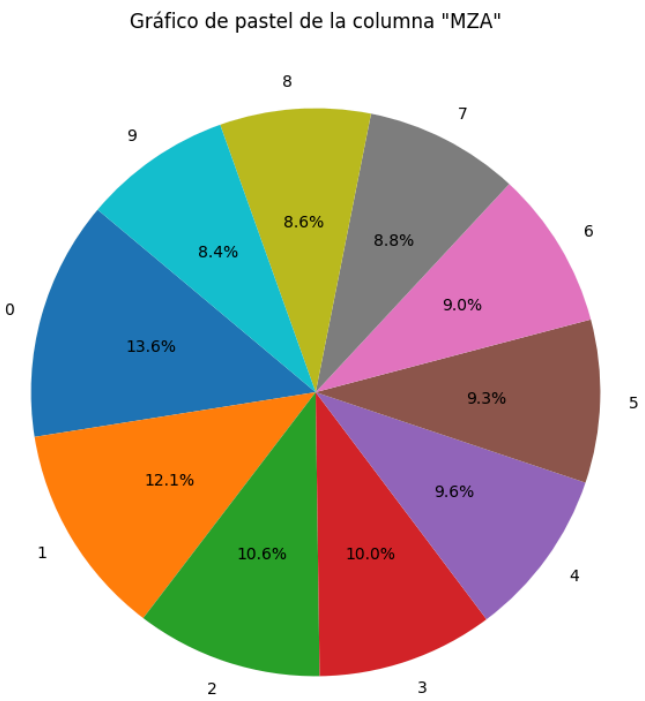


Figura 23. Gráfico de pastel de la columna ‘MZA’.

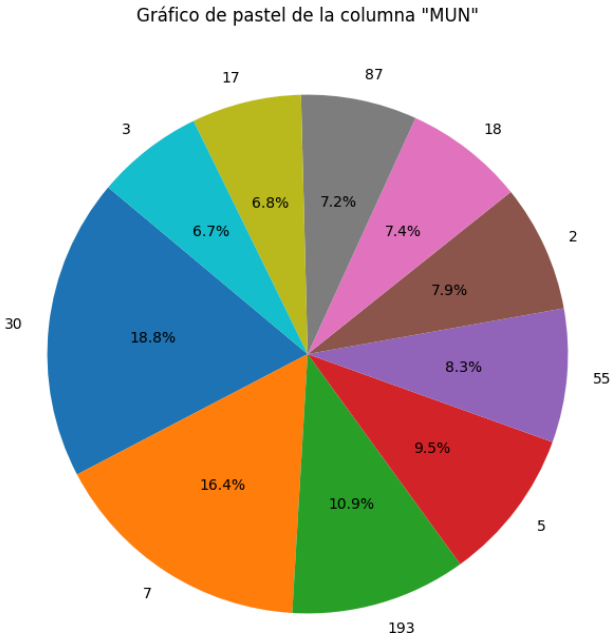


Figura 24. Gráfico de pastel de la columna ‘MUN’.

Verifica la calidad de los datos: ¿existen valores faltantes, valores incorrectos en los datos, errores de ortografía?

En el dataset encontramos valores * y valores N/A los cuales cambiamos por np.NAN los cuales eliminamos con dropna Esto con el objetivo de tener una base de datos limpia y mejorar el procesamiento de los datos. Luego, cambiamos el tipo de datos de objeto a flotante ya que a pesar de ser numéricos los datos, el dataset los leía como objeto.

Genera Modelo de Datos en UML o ER a partir de la estructura del archivo csv (fuente de datos)

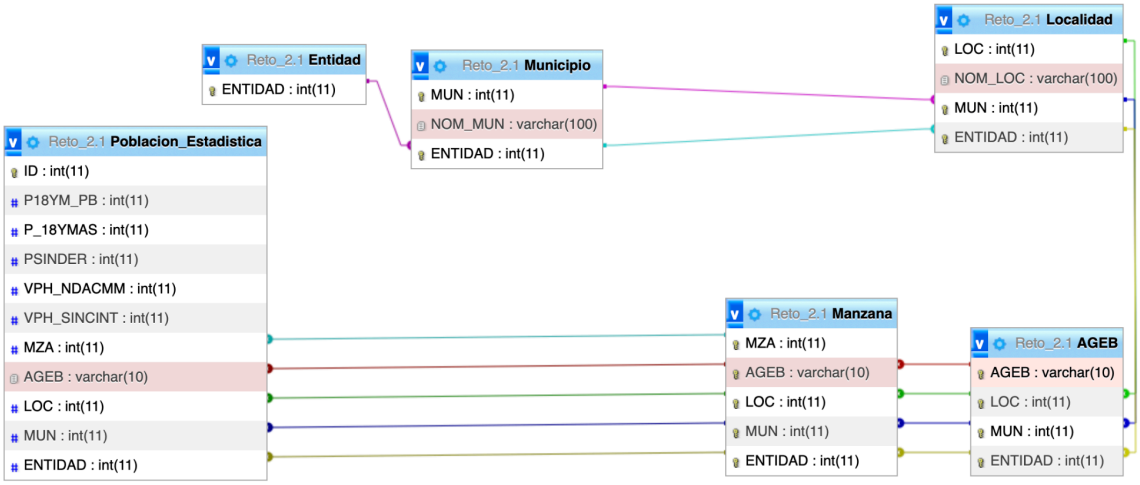


Figura 25. Diagrama ER.

2) Preparación de los datos:

1. Selecciona el conjunto de datos a utilizar de acuerdo al objetivo del proyecto.

a) Decide qué conjunto de datos se utilizará. Explica por qué se incluyeron o excluyeron ciertos datos.

En la segunda etapa, se basó más en seleccionar las columnas más importantes del dataset que son las localidades, municipios, manzanas, ageb, población total de arriba de 18 años, población mayor de 18 años con educación postbásica terminada, población sin afiliación a servicios de salud, viviendas particulares que no disponen de automóvil o camioneta, de motocicleta o motoneta y viviendas particulares habitadas sin computadora ni Internet.

b) Identifica columna objetivo, columna que se desea predecir.
Las columnas 'VPH_SINCINT', 'PSINDER', 'VPH_NDACMM' las utilizaremos para predecir la columna ‘P_18YM_PB’ que es la población de 18 años y más con educación posbásica.

2.- Limpieza de datos
a) Elimina duplicados
b) Corrige valores erróneos.
c) Maneja valores faltantes.
d) Maneja datos categóricos: Transforma datos categóricos a datos numéricos si es necesario.
e) Maneja adecuadamente los valores atípicos (outliers) que encuentres en el dataset.

3.- Transformación de Datos:

a) Revisa si es necesario discretizar los datos (binning)

b) Si es necesario escala y normaliza los datos.

c) Si lo consideras conveniente, construye nuevos atributos que se puedan obtener a partir de los datos disponibles. (atributos derivados).

4.- Reformatea/reestructura los datos si es necesario
Se eliminaron los valores nulos , por lo tanto no hicimos estos pasos como tal.

Autoevaluación

Nombre	**Evaluación**
Marcos Aquino Garcia A00835576	**10**
Sergio Alejandro Esparza González A01625430	**10**
Carlos Alberto Gómez San Pedro A01658377	**10**
Pedro Soto Juárez A00837560	**10**
José Francisco Obregón Gaxiola A00227502	**10**
Arath Mendivil Mora A01660670	**10**

Etapa 3

1) Documentación mejorada de fases anteriores

Realizar cambios en la documentación de las fases anteriores de acuerdo a las sugerencias obtenidas.

2) Fase de Generación de Modelos de aprendizaje automático.

Incluir lo siguiente:

a) Variables relevantes para el análisis

Mostrar en una tabla las variables relevantes para el proceso de modelación de algoritmos de aprendizaje. Incluir para cada variable su nombre, descripción, tipo de dato (entero/string) y tipo de variable (categórica nominal/ordinal, numérica. Indicar en la tabla cuáles son las variables predictoras y cuál es la variable objetivo o resultado.

```
ENTIDAD        int64
MUN            int64
NOM_MUN       object
LOC            int64
NOM_LOC       object
AGEB          object
MZA            int64
P18YM_PB      object
P_18YMAS      object
PSINDER       object
VPH_NDACMM    object
VPH_SINCINT   object
dtype: object
```

```
#Cambiamos las variables object que eran números a float
dfc3['P18YM_PB'] = dfc3['P18YM_PB'].astype(float)
dfc3['P_18YMAS'] = dfc3['P_18YMAS'].astype(float)
dfc3['LOC'] = dfc3['LOC'].astype(float)
dfc3['MUN'] = dfc3['MUN'].astype(float)
dfc3['MZA'] = dfc3['MZA'].astype(float)
dfc3['VPH_SINCINT'] = dfc3['VPH_SINCINT'].astype(float)
dfc3['PSINDER'] = dfc3['PSINDER'].astype(float)
dfc3['VPH_NDACMM'] = dfc3['VPH_NDACMM'].astype(float)
```

Nombre	Descripción	Tipo de dato	Tipo de variable	Tipo
PSINDER	Población sin afiliación a servicios de salud.	Float	Numérico.	Predictora

VPH_NDACMM	Viviendas particulares que no disponen de automóvil o camioneta, de motocicleta o motoneta	Float	Numérico	Predictora
VPH_SINCINT	Viviendas particulares habitadas sin computadora ni Internet.	Float	Numérico	Predictora
P18YM_PB	Población de 18 años y más con educación posbásica	Float	Numérico	Objetivo

P18YM_PB: Población de 18 años y más con educación posbásica. Es numérico. Puede tomar cualquier valor entero arriba de 0.
0.
P_18YMAS: Población de 18 años y más. Es numérico. Puede tomar cualquier valor entero arriba de 0.
PSINDER: Población sin afiliación a servicios de salud. Es numérico. Puede tomar cualquier valor entero de 0 a infinito.
VPH_NDACMM: Viviendas particulares que no disponen de automóvil o camioneta, de motocicleta o motoneta. Puede tomar cualquier valor entero. Puede tomar valor de 0 a infinito.
VPH_SINCINT: Viviendas particulares habitadas sin computadora ni Internet. Puede tomar cualquier valor entero. Puede tomar valor de 0 a infinito.

b) Datos disponibles

1) Mostrar una imagen de la tabla de datos, con al menos 10 registros, considerando sólo las variables/columnas relevantes para esta fase.

Código + Texto

Disco Colab AI

```
dfc4=dfc3[['P18YM_PB','PSINDER','VPH_NDACMM','VPH_SINCINT']]
dfc4.head(11)
```

	P18YM_PB	PSINDER	VPH_NDACMM	VPH_SINCINT
0	4516388.0	2502789.0	1385629.0	561128.0
1	234812.0	90370.0	65802.0	22687.0
2	234812.0	90370.0	65802.0	22687.0
3	1751.0	603.0	376.0	148.0
4	97.0	27.0	25.0	9.0
5	76.0	46.0	15.0	9.0
6	64.0	17.0	16.0	6.0
7	89.0	38.0	14.0	7.0
8	93.0	21.0	17.0	7.0
9	81.0	23.0	13.0	4.0
10	103.0	31.0	22.0	8.0

2) Mostrar el total de datos disponibles (filas y columnas) y la distribución de los mismos de acuerdo a la variable objetivo. ¿Se tienen datos balanceados?

Disco

```
filas,columnas=dfc4.shape
print("Este es el número de filas:",filas )
print("Este es el número de columnas:",columnas )
```

```
Este es el número de filas: 155527
Este es el número de columnas: 4
```

0 s

```
[32] # Distribución de los datos según la variable objetivo
distribucion_variable_objetivo =df_cuan['P18YM_PB'].value_counts()
print("\nDistribución de los datos según la variable objetivo:")
print(distribucion_variable_objetivo)
```

```
Distribución de los datos según la variable objetivo:
P18YM_PB
0.0         20745
7.0          3180
9.0          3143
10.0         3127
8.0          3123
            ...
1577.0          1
1712.0          1
```

0 s completado a las 10:25

```
[32] # Distribución de los datos según la variable objetivo
     distribucion_variable_objetivo =df_cuan['P18YM_PB'].value_counts()
     print("\nDistribución de los datos según la variable objetivo:")
     print(distribucion_variable_objetivo)
```

```
Distribución de los datos según la variable objetivo:
P18YM_PB
0.0         20745
7.0          3180
9.0          3143
10.0         3127
8.0          3123
            ...
1577.0          1
1712.0          1
1534.0          1
309959.0        1
7540.0          1
```

✓ 0 s completado a las 10:25

```
# Verificación de balance de datos
balanceado = distribucion_variable_objetivo.min() / distribucion_variable_objetivo.max() > 0.8  # Se considera b
print("\n¿Los datos están balanceados?:", balanceado)
```

```
¿Los datos están balanceados?: False
```

```
[ ] #Usamos describe para calcular los valores estadisticos de las variables cuantitativas
```

En un modelo de regresión que vamos a usar en un random forest no importa que los datos esten desbalanceados.

c) Modelos de Aprendizaje Supervisado

1) Elaborar un resumen de modelos de clasificación, modelos de regresión y la diferencia entre ellos. ¿Qué tipo de modelos se aplicarán, clasificación o regresión? ¿Por qué?

2) Describir brevemente los modelos de aprendizaje automático que se utilizarán, incluyendo los principales parámetros que se definen y ventajas/desventajas de cada uno

1. Decision Trees

Parámetros Principales:

- max_depth: Profundidad máxima del árbol.
- min_samples_split: Número mínimo de muestras necesarias para dividir un nodo.
- min_samples_leaf: Número mínimo de muestras requeridas en un nodo hoja.

Ventajas:

- Fácil de entender e interpretar: Los árboles pueden ser visualizados.
- Requiere poca preparación de datos: No necesita normalización de datos.
- Puede manejar datos tanto numéricos como categóricos.

Desventajas:

- Sobreajuste: Tienden a memorizar los datos y sobre ajustarse, especialmente si el árbol es muy profundo.
- No muy robustos: Un pequeño cambio en los datos puede llevar a un árbol muy diferente.

2. Redes Neuronales
Parámetros Principales:

- Número de capas y número de neuronas por capa: Define la estructura de la red.
- Función de activación: Determina la transformación no lineal a aplicar (p.ej., ReLU, sigmoid).
- Ratio de aprendizaje: Velocidad con la que la red actualiza sus pesos durante el entrenamiento.

Ventajas:

- Capacidad de aprendizaje de representaciones complejas: Pueden modelar relaciones no lineales complejas.
- Buen rendimiento en grandes conjuntos de datos.

Desventajas:

- Requieren gran cantidad de datos para entrenarse efectivamente.
- Costosos computacionalmente.
- Difíciles de interpretar: Las redes neuronales son a menudo consideradas como "cajas negras".

3. K-Nearest Neighbors (KNN)
Parámetros Principales:

- n_neighbors (K): Número de vecinos a considerar.
- weights: Peso de los vecinos, puede ser uniforme o variar según la distancia.
- metric: Métrica de distancia para medir la cercanía entre puntos.

Ventajas:

- Simple de entender e implementar.
- No necesita entrenamiento explícito, lo que puede ser ventajoso en ciertos escenarios.

Desventajas:

- Ineficiente para grandes conjuntos de datos: La necesidad de almacenar y hacer cálculos con todo el conjunto de entrenamiento puede ser computacionalmente

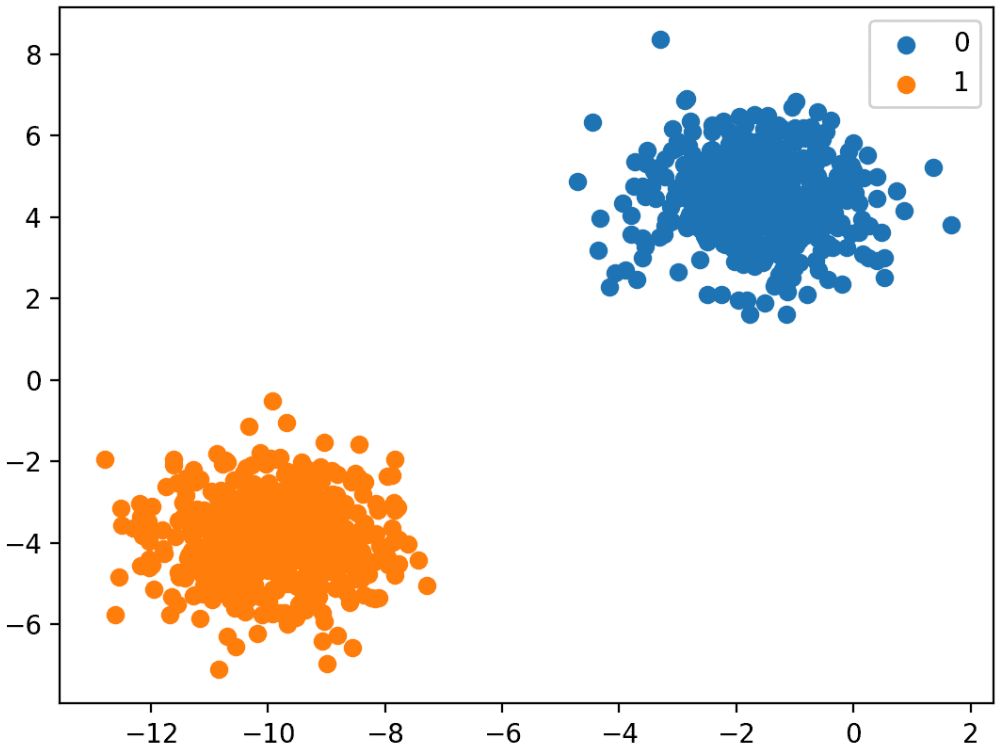
costosa.

- Sensible a características irrelevantes y a la escala de los datos.

Clasificación se refiere a predecir una cierta categoría o clase donde se predice , dada una cierta cantidad de entrada de datos, de una perspectiva de entrenamiento de datos se requiere de muchas entradas de datos y de salidas para que un modelo aprenda, un ejemplo muy conocido es saber si un correo es spam o no, el accuracy o exactitud de clasificación es una métrica para saber que tan bien rinde un modelo basado en sus predicciones de categorías.

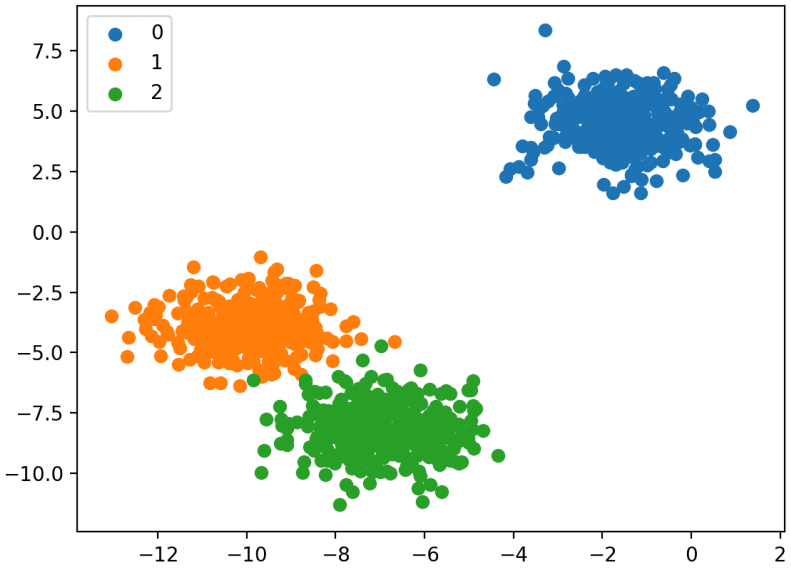
Hay 4 tipos de principales clasificadores que son Binaria,Multiclase,Multi-Label y desequilibrada

Los de clasificación binaria que involucra dos clases , una que es su estado normal y otra que es el estado abnormal , supongamos que el estado normal es cáncer no detectado y por otro lado el abnormal es ahora ya el cáncer detectado.No soportan más de 2 clases , se usan estos algoritmos : regresión logística ,k nearest neighbors,decision trees,support vector machine y naive bayes.



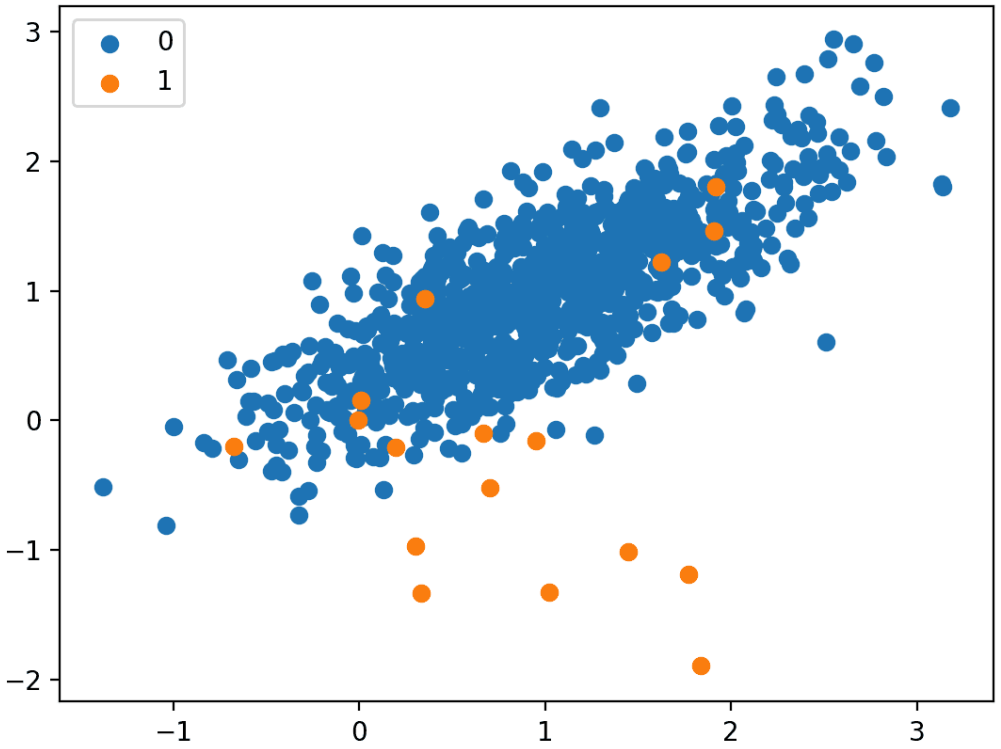
Los de multiclase se refiere cuando tiene que clasificar cuando hay más de dos etiquetas , una de las aplicaciones sería reconocer la cara de alguien entre varias fotos que hay de varias personas, usan estos algoritmos: regresión logística ,k nearest neighbors,decision

trees,support vector machine , naive bayes y gradient boosting.



Los de multi label o multi etiqueta , es lo mismo que el anterior , difiere que en vez de predecir solo una clase , entre varias ,puede predecir o detectar varias cosas al mismo tiempo, en una foto , el poder reconocer a un perro,gato , una manzana,etc. Se usan estos algoritmos:
Multi-label Decision Trees
Multi-label Random Forests
Multi-label Gradient Boosting

En clasificación no balanceada son tareas de clasificación donde el número de ejemplos de las clases están desbalanceadas , esto incluye detección de fraudes, diagnósticos médicos ,etc.
Se usan estos algoritmos:
Cost-sensitive Logistic Regression.
Cost-sensitive Decision Trees.
Cost-sensitive Support Vector Machines.



Esos son algunos tipos de modelos de clasificación , ahora se explica los más importantes de regresión y los que usamos.

La regresión se utiliza en lugar de la clasificación cuando el objetivo principal es predecir un valor numérico en vez de asignar etiquetas o clases , que es lo que más se adecua a nuestro proyecto.

La regresión lineal se usa para análisis predictivo , es la forma más básica de análisis de regresión se utiliza para modelar una relación lineal entre una variable dependiente y una o más variables independientes.

Una regresión de árbol de decisión es una herramienta para clasificación y para predicción de valores continuo ,donde un árbol de decisión es un tipo de diagrama de flujo donde iterativamente se crean subconjuntos de datos teniendo condiciones.

El random forest , es un algoritmo que combina varios árboles de decisión para predecir valores continuos.

En K-Nearest Neighbors , la predicción se realiza calculando el promedio o mediana de los valores de las k instancias más cercanas al punto de prueba. En lugar de votar por una clase como en la clasificación, los valores de las k instancias más cercanas se promedian para estimar el valor continuo de la variable objetivo.

Ridge regulariza el modelo resultante imponiendo una penalización al tamaño de los coeficientes de la relación lineal entre las características predictivas y la variable objetivo. En este caso, los coeficientes calculados minimizan la suma de los cuadrados de los residuos penalizada al añadir el cuadrado de la norma L2 del vector formado por los coeficientes.

Lasso regression es un modelo lineal que penaliza el vector de coeficientes añadiendo su norma L1 (basada en la distancia Manhattan) a la función de coste:

Lasso tiende a generar "coeficientes dispersos": vectores de coeficientes en los que la mayoría de ellos toman el valor cero. Esto quiere decir que el modelo va a ignorar algunas de las características predictivas.
Esto hace que el modelo sea más sencillo de interpretar ya que puede poner de manifiesto las características más importantes del conjunto de datos.
En el caso de que exista cierta correlación entre las características predictivas, Lasso tenderá a escoger una de ellas al azar.

Elasticnet regression
Combina las penalizaciones L1 y L2 de Ridge y Lasso . Esto proporciona una mayor flexibilidad al permitir la selección automática de características y la reducción de multicolinealidad. ElasticNet es útil en conjuntos de datos con características correlacionadas.

Usamos KNN neighbors,Random Forest y árboles de decisión ya que fueron los que más se acoplaron a nuestro proyecto y tuvieron mejores resultados usando regresión usando gridsearchcv que es una técnica de validación cruzada para encontrar los mejores parámetros.
No hicimos algoritmos de clasificación.

Material de apoyo:
https://machinelearningmastery.com/types-of-classification-in-machine-learning/Links to an external site.

https://machinelearningmastery.com/multinomial-logistic-regression-with-python/Links to an external site.

d) Generación y Evaluación de los Modelos de Aprendizaje

1.- Por cada miembro del equipo, incluir imágenes de la aplicación de modelos de aprendizaje automático y los resultados obtenidos

2.- Considerando los resultados obtenidos por cada miembro del equipo al generar y evaluar, individualmente, modelos de aprendizaje en su propio notebook en Python, generar una tabla incluyendo, para cada modelo aplicado, el tipo de modelo, valor de parámetros utilizados (diferentes entre los miembros del equipo) y resultados de evaluación de la métrica de exactitud (accuracy). La tabla debe seguir el siguiente formato, dependiendo del número de alumnos por equipo.

2.- Señalar en la tabla, para cada tipo de modelo, el mejor resultado obtenido.
3.- Análisis de resultados y selección del modelo de predicción.

Tipo de Modelo	Pedro Soto Juárez	Francisco Obregón Gaxiola	Sergio Alejandro Esparza González	Marcos Renato Aquino	Arath Mendivil Mora	Carlos San Pedro
Random Forest	Parámetros utilizados: max_depth:20,min_samples_split:2,n_estimators:20 Resultados o: Coeficiente de determinación (R^2): 0.9012	Parámetros utilizados: max_depth: 20, min_samples_split: 2, n_estimator: 5 Resultados obtenidos: Coeficiente de determinación (R^2): 0.9881	Parámetros utilizados: max_depth: 20, min_samples_split: 2, n_estimator: 20 Resultados obtenidos: Coeficiente de determinación (R^2): 0.8107	Parámetros utilizados: max_depth:10,min_samples_split:2,n_estimators:20 Resultados obtenidos: Coeficiente de determinación (R^2): 0.98449	Parámetros utilizados: max_depth:10,min_samples_split:2,n_estimators: 20 Resultados obtenidos: Coeficiente de determinación (R^2): 0.7596	Parámetros utilizados: max_depth:10,min_samples_split:2,n_estimators:20 Resultados obtenidos: Coeficiente de determinación (R^2): 0.9958

K Neighbors	Parámetros utilizados: k=2 Resultados obtenidos: Coeficiente de determinación (R^2) con el mejor modelo: 0.4384	Parámetros utilizados: k=2 Resultados obtenidos: Coeficiente de determinación (R^2) con el mejor modelo: 0.4384	Parámetros utilizados: k=4 Resultados obtenidos: Coeficiente de determinación (R^2) con el mejor modelo: 0.8200	Parámetros utilizados: k=2 Resultados obtenidos: Coeficiente de determinación (R^2) con el mejor modelo: 0.4384	Parámetros utilizados: k=2 Resultados obtenidos: Coeficiente de determinación (R^2) con el mejor modelo:0.7595	Parámetros utilizados: k=2 Resultados obtenidos: Coeficiente de determinación (R^2) con el mejor modelo: 0.4384
Decision Trees	Parámetros utilizados: Max depth:10 y min_samples_split:2 Resultados obtenidos:0.4893 de coeficiente de determinación(R^2)	Parámetros utilizados: Parámetros utilizados: Max depth:20 y min_samples_split:2 Resultados obtenidos: 0.4911 de coeficiente de determinación(R^2)	Parámetros utilizados: max_depth: 10, min_samples_split: 2 Resultados obtenidos: Coeficiente de determinación (R^2): 0.7125	Parámetros utilizados: Max depth:5 y min_samples_split:2 Resultados obtenidos :0.4910 de coeficiente de determinación(R^2)	Parámetros utilizados: Max depth:20 y min_samples_split: 2 Resultados obtenidos : 0.7596	Parámetros utilizados: Max depth:10 y min_samples_split:2 Resultados obtenidos: 0.4910 de coeficiente de determinación(R^2)

Tipo de Modelo	Pedro Soto Juárez	Francisco Obregón Gaxiola	Sergio Alejandro Esparza González	Marcos Renato Aquino	Arath Mendivil Mora	Carlos San Pedro
Random Forest	max_depth:20,min_samples_split:2,n_estimators:20 Coeficiente de determinación (R^2): 0.9012	max_depth: 20, min_samples_split: 2, n_estimator: 5 (R^2): 0.9881	max_depth: 20, min_samples_split: 2, n_estimator: 20 0.8107	max_depth:10,min_samples_split:2,n_estimators:20 (R^2): 0.98449	max_depth:10,min_samples_split:2,n_estimators: 20 (R^2): 0.7596	max_depth:10,min_samples_split:2,n_estimators:20 (R^2): 0.9958
K Neighbors	: k=2 (R^2): 0.4384	: k=2 (R^2): 0.4384	k=4 (R^2): 0.8200	k=2 (R^2: 0.4384	k=2 (R^2):0.7595	k=2 (R^2): 0.4384
Decision Trees	Max depth:10 y min_samples_split:2 0.4893 de coeficiente de determinación(R^2)	Max depth:20 y min_samples_split:2 0.4911	Max_depth: 10, min_samples_split: 2 0.7125	Max depth:5 y min_samples_split:2 0.4910	Max depth:20 y min_samples_split: 2 0.7596	Max depth:10 y min_samples_split:2 0.4910

e) Conclusiones de esta etapa.

f) Anexo - Individual

Describir las actividades realizadas y el aprendizaje obtenido.

4) Selección y Despliegue

Selección del modelo considerando las métricas más apropiadas de acuerdo a los criterios de éxito del negocio.

Elegimos Random Forest para en regresión

Ejemplo de aplicación del modelo
Descripción del prototipo funcional

Optimización de recursos: Utilizar el modelo para identificar áreas geográficas o grupos de población con una mayor propensión a no completar la educación post-básica, lo que permitiría dirigir los recursos y las intervenciones hacia donde sean más necesarios.

5) Conclusiones y Recomendaciones

Interpretación de los resultados obtenidos, su impacto y utilidad.
Sumarizar todo el proceso, principales problemas.
Recomendaciones al socio formador
Siguientes pasos que se recomienda hacer

Conclusión

El estudio confirmó la viabilidad de usar modelos de regresión para predecir la educación post básica y las condiciones de vida asociadas. El Random Forest, con su capacidad para manejar múltiples variables y capturar complejidades no lineales, se estableció como el modelo más prometedor. Gracias a esto, se puede implementar en distintos municipios una forma de optimizar los recursos para que así más jóvenes mexicanos logren terminar sus estudios de preparatoria-

Actividades por realizar

Monitorear y actualizar el modelo regularmente con nuevos conjuntos de datos para mantener su relevancia.
Identificar áreas geográficas o grupos de población con una mayor propensión a no completar la educación post-básica.

Recomendaciones a INEGI

Promover este modelo en los distintos Estados de México para así optimizar el manejo de recursos donde es más necesario.

Conclusiones:

En esta fase del proyecto, identificamos las variables relevantes para nuestro análisis. Utilizamos cuatro variables específicas: la población sin afiliación a servicios de salud (PSINDER), el número de viviendas particulares sin vehículos (VPH_NDACMM), las viviendas sin acceso a computadora e Internet (VPH_SINCINT), y la población de 18 años y más con educación posbásica (P18YM_PB) que sería la variable objetivo o la que queremos predecir con los modelos. Estas variables fueron seleccionadas debido a su importancia para nuestro objetivo de predecir el nivel de educación posbásica en tres estados de la república. A través de modelos de aprendizaje automático, buscamos comprender cómo estas variables podrían influir en dicho nivel educativo.

Después de aplicar diversos modelos de aprendizaje automático, hemos obtenido resultados variados en términos de su capacidad para predecir el nivel de educación posbásica en los tres estados de la república. En general, observamos que el modelo de Random Forest obtuvo el mejor desempeño, con un coeficiente de determinación (R^2) cercano a 1, lo que indica una buena capacidad predictiva. No obstante, también notamos que los modelos de Decision Trees y K-Nearest Neighbors tuvieron un desempeño inferior en comparación con Random Forest. Estos resultados nos dicen que las variables predictoras que seleccionamos pudieron tener un impacto significativo en la predicción del nivel educativo posbásico, y que el modelo de Random Forest es el más adecuado para este conjunto de datos y este problema en específico. Estamos satisfechos con lo aprendido durante la etapa de generación y evaluación de modelos de aprendizaje automático. Hemos trabajado y explotado una variedad de algoritmos y técnicas para predecir el nivel de educación posbásica en tres estados de la república, lo que nos ha permitido entender mejor la relación entre las variables predictoras y la variable objetivo.

Anexo:

Pedro Soto Juarez:

Artículos consultados:
Causas y consecuencias de la deserción escolar en el bachillerato:Caso universidad autónoma de sinaloa
Resumen: Una de las causas que más influyen en la deserción de los estudios de preparatoria son: casarse y el reprobar materias , donde se destaca en este estudio que las mujeres desertan más que los hombres debido a que quieren salir de su cotidianidad y buscar otras oportunidades al lado de su novio sin embargo esto causa que tengan que atender una familia a una temprana edad y se les priva de un desarrollo profesional , la segunda causa es el que no están motivados los estudiantes por lo tanto no les gusta estudiar , el factor económico de que sus padres no ganan de una manera que puedan seguir pagandoles los estudios y la falta de apoyo de los padres que estudien.

Predicting student dropout in Self-Paced MOOC course using Random Forest model y Early prediction of university dropouts – a random forest approach.
Se trata de buscar si una persona va a seguir estudiando la universidad o en un MOOC que son cursos en línea esto se basa en inputs de cómo se sienten con las clases , su estado del ánimo, cuánto tiempo estudian , entre otros , y se usa el random forest para clasificar si es que van a abandonar los estudios o no.

Se llevó a cabo cambios en el objetivo de proyecto debido a que faltaban variables predictoras y enfocarnos en una variable objetivo que es la población que terminó la educación postbásica en tres estados de la república , utilizando 3 modelos de regresión lineal donde primeramente teníamos que separar datos de prueba y entrenamiento para implementar los distintos modelos que son árboles de decisión , random forest y KNN neighbors ,antes de implementarlos usamos gridsearchcv para encontrar hiperparámetros de tal manera que nos daban valores muy bajos de coeficiente de determinación en decision trees , knn neighbors , el único decente fue en mi caso que salía 0.90 en Random forest que quiere decir que las variables predictoras que son las personas con falta de Internet , personas sin afiliación a un servicio de salud y personas sin transporte si tienen un gran impacto en terminar la educación post-básica o también ocurre en factores indirecto tal es como la falta de transporte se relaciona con que viven en zonas rurales y hay falta de disponibilidad de recursos.

Francisco Obregon:

En esta última fase, se identificaron las variables relevantes para el análisis, convirtiendo los datos necesarios a formatos adecuados y describiendo cada variable en detalle. Para la selección de modelos de aprendizaje supervisado, optamos por utilizar algoritmos como Random Forest, K-Nearest Neighbors y Decision Trees. Cada uno de estos modelos tiene sus

propias ventajas y desventajas, y elegimos aquellos que se adaptan mejor a las necesidades específicas de predicción numérica. En la fase de generación y evaluación de modelos, nosotros como equipo aplicamos diferentes modelos de aprendizaje automático en sus notebooks de Python y evaluamos los resultados obtenidos. Después recopilamos los resultados en una tabla comparativa, destacando el mejor resultado obtenido para cada tipo de modelo con los parámetros especificados. El proceso nos permitió aprender sobre la implementación práctica de los diferentes algoritmos de aprendizaje automático, así como sobre la importancia de ajustar correctamente los parámetros de cada modelo. También adquirimos experiencia en la evaluación de la precisión de los modelos y en la selección del mejor modelo para nuestras necesidades específicas de predicción.

Sergio Alejandro Esparza González:

En este proyecto acerca de la educación posbásica, se llevaron a cabo varios análisis que tuvieran impacto para su realización, por ejemplo, la selección de variables para su posterior limpieza y preparación para el análisis utilizando notebooks de Google Colaboratory con el lenguaje de Python. Después se ajustaron distintos modelos de aprendizaje automático utilizando distintos algoritmos como K-Nearest Neighbors, Decision Trees y Random Forests. Con esta experiencia, pudimos validar las hipótesis iniciales para los factores del éxito educativo usando modelación estadística y ciencia de datos. Se tomaron métricas de desempeño para evaluar el mejor algoritmo y se hicieron comparaciones acerca de los resultados para tomar las mejores decisiones posibles.

Carlos Alberto Gómez San Pedro:
En la etapa final del análisis de nuestro proyecto sobre la educación posbásica, nos enfocamos en refinar los modelos de predicción empleados para entender mejor los factores que contribuyen al éxito educativo. Tras una evaluación inicial, decidimos ajustar los modelos de Random Forest, K-Nearest Neighbors y Decision Trees para adaptarlos específicamente a las peculiaridades de nuestro dataset. En esta fase, se dividió el dataset en conjuntos de entrenamiento y prueba para evaluar la efectividad de cada modelo en condiciones controladas. Durante la implementación, empleamos técnicas de optimización de hiper parámetros como Grid Search CV para cada modelo, lo que resultó en una mejora notable en la precisión de las predicciones. Random Forest continuó mostrando un desempeño superior, con un coeficiente de determinación significativamente alto, resaltando su capacidad para manejar la heterogeneidad de los datos socioeconómicos asociados a la educación postbásica. Esta experiencia nos permitió no solo validar nuestras hipótesis iniciales sobre los factores críticos para el éxito educativo, sino también mejorar nuestra competencia en técnicas avanzadas de modelado estadístico.

Arath Mendivil Mora:

Durante la fase de consolidación de datos para nuestro proyecto, se llevó a cabo una selección meticulosa de las variables más impactantes para la educación postbásica. Se realizó una limpieza y normalización de datos para prepararlos para el análisis, asegurando que los datos estuvieran completos y listos para su procesamiento en modelos estadísticos. Los modelos de aprendizaje automático seleccionados para este estudio fueron Random Forest, K-Nearest Neighbors y Decision Trees, por su relevancia en la predicción de fenómenos sociales complejos. Implementamos estos modelos en nuestros entornos de desarrollo utilizando Python, aplicando técnicas de validación cruzada para optimizar los hiper parámetros y mejorar la fiabilidad de las predicciones. Los resultados obtenidos de Random Forest fueron particularmente destacados, mostrando un alto coeficiente de determinación, lo que sugiere una fuerte correlación entre las variables predictoras y el éxito en la educación postbásica. Esta etapa del proyecto no solo refinó nuestro enfoque analítico, sino que también profundizó nuestro entendimiento del impacto de las condiciones socioeconómicas en los resultados educativos.

Marcos Renato Aquino:

En esta etapa, una vez con los datos que se iban a utilizar definidos y limpiados, definimos las variables X que nos iban a ayudar a predecir la variable y. Para esto, probamos cada uno con tres distintos tipos de modelos de regresión ya que la variable que necesitábamos es un valor numérico. Las tres que utilizamos fueron Decision Trees, KNN neighbors y random forest en las cuales el random forest nos daba un R2 preciso y constante cercano a uno, lo que indica que es el modelo indicado para poder cumplir con nuestro objetivo. Con esto logramos perfeccionar distintos tipos de aprendizaje supervisado y cómo aplicarlo en casos reales.

Referencias bibliográficas

[1] Behr, A., Giese, M., K, H. D. T., & Theune, K. (2020). Early prediction of university dropouts – a random forest approach. Jahrbücher Für Nationalökonomie Und Statistik, 240(6), 743–789. https://doi.org/10.1515/jbnst-2019-0006

[2] Castro, M. (2023, 9 mayo). El abandono escolar también tiene género. IMCO. https://imco.org.mx/el-abandono-escolar-tambien-tiene-genero/

[3] Christian, M. (2015). School Location, School Section and Students' Gender as Predictors to Secondary School Dropout Rate in Rivers State, Nigeria. Journal of Education and Practice, 6(28), 113–118.

[4] Dass, S., Gary, K., & Cunningham, J. (2021). Predicting student dropout in Self-Paced MOOC course using Random Forest model. Information, 12(11), 476. https://doi.org/10.3390/info12110476

[5] De Educación Pública, S. (s. f.). Boletín 100 Desciende a 8.1% tasa de abandono escolar en Educación. https://www.gob.mx/sep/articulos/boletin-100-desciende-8-1-tasa-de-abandono-escolar-en-educacion-superior-sep?idiom=es#:~:text=El%20Gobierno%20de%20M%C3%A9xico%2C%20por,inform%C3%B3%20el%20subsecretario%20de%20ese

[6] González, M. T. G. (2015). Los centros escolares y su contribución a paliar el desenganche y abandono escolar. Profesorado. Revista de Curriculum y Formación de Profesorado, 19(3), 158-176.

[7] Red por los Derechos de la Infancia en México. (4 de diciembre de 2023). Abandono escolar de niñas, niños y adolescentes en México (2016-2021). Infancia Cuenta en México. Recuperado de https://blog.derechosinfancia.org.mx/2023/12/04/abandono-escolar-de-ninas-ninos-y-adolescentes-en-mexico-2016-2022/

[8] Ruiz-Ramírez, R., García-Cué, J. L., & Pérez-Olvera, M. A. (2014). CAUSAS Y CONSECUENCIAS DE LA DESERCIÓN ESCOLAR EN EL BACHILLERATO: CASO UNIVERSIDAD AUTONÓMA DE SINALOA. Ra Ximhai, 10(5), 51-74.